Predicting Controversiality of Films

5/14/2020

Introduction

Movie-making is a major industry in the United States, comprising nearly 3.2% of US good and services in 2011,[1] with its related culture dominating news cycles and social media topics across the globe. Yet the industry is changing, as digital disruptors pull audiences away from theaters and to subscription streaming services.[2] With an increasingly crowded industry, with content being produced and released through non-traditional channels, it is becoming increasingly difficult for individual films to stand out from their peers and produce the same impact on society that major cinematic blockbusters traditionally pioneered. For this project I take a look into historical release of films, and search for patterns existing within movie data.

Data Source

For this project I will be using 'The Movies Dataset', a collection of metadata on 45,000 movies, with 26 million ratings compiled on Kaggle.[3] This database was collected from two primary sources: IMDb and MovieLens.

IMDb purports to be the "world's most popular and authoritative source for information on movies, TV shows, and celebrities".[4] Sourced from IMDb, is metadata for films which includes box office revenues, budgets, cast information, and plot summaries - among other features. It includes movies released as far back as 1874 through July 2017.

MovieLens is a research site designed to create personalized movie recommendations.[5] From MovieLens, the data set contains ratings ranging from 0.5 to 5 for films from over 270,000 users. Additionally, MovieLens supplied keywords for each film that clue to the content, or story, of each film.

[1] Hollywood has blockbuster impact on US economy that tourism fails to match. (2013, December 5). Retrieved May 11, 2020, from https://www.theguardian.com/business/2013/dec/05/arts-culture-us-economy-gdp

[2] Hurtz, B. (2017, May 25). Netflix, Hollywood's biggest disruptor, is radically altering the movie landscape. Retrieved May 13, 2020, from https://www.theglobeandmail.com/arts/film/netflix-hollywoods-biggest-disruptor-is-radically-altering-the-movielandscape/article35115001/

[3] https://www.kaggle.com/rounakbanik/the-movies-dataset

[4] Press Room. (n.d.). Retrieved from https://www.imdb.com/pressroom/about/

[5] About MovieLens. (n.d.). Retrieved May 11, 2020, from https://movielens.org/info/about

To appropriately work with this data, I constructed a database designed to join the IMDb metadata to the MovieLens ratings using common identification numbers, provided by the Links file on Kaggle.

Prediction Target

There is value in producing films that provoke strong emotional feelings. Strong emotions invoked by a film's contents can drive broad discussions both online, and in the news media - increasing awareness and interest in the film, which can act to ticket sales or online streams. From this dataset, I will examine information related to the contents of a film to see if it may be used to predict this kind of controversiality that can drive virality.

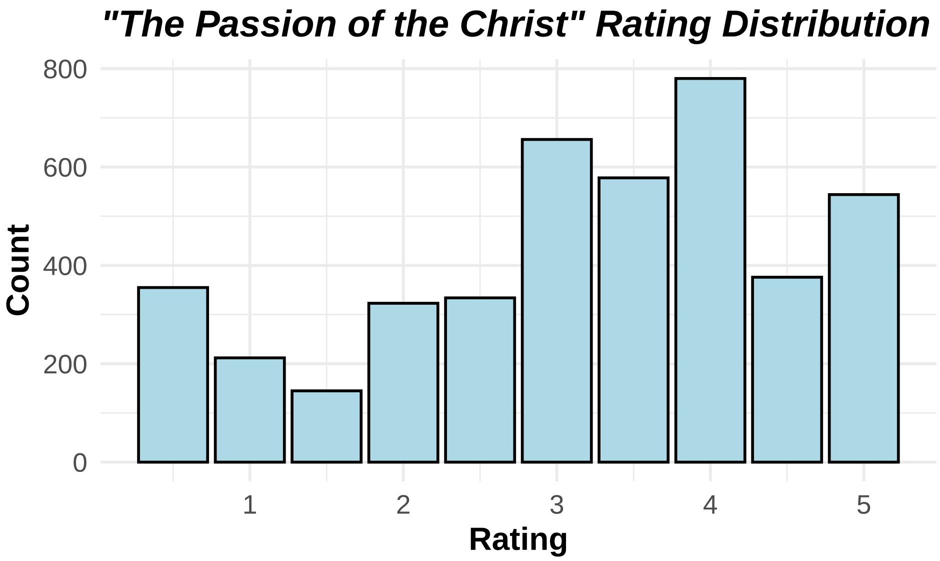
Defining the Target

Since there is no tag for controversiality inherent to the database, I constructed an artificial feature relying the ratings information from MovieLens. I believe that the strong feelings would be reflected in the ratings. For purposes of this analysis I define the term "polarized ratings" to be ratings of 0.5 or 1.0 stars (negative polarization) and scores of 4.5 or 5.0 stars (positive polarization).

I expect controversial films would have a high proportion of both negative and positive polarized ratings. Codifying it into the construction of a Controversial tag, I determined that Controversial movies would have polarized ratings construing at least 30% of all ratings.

To ensure balance between negative and positive polar ratings, I defined and included a Polarity Ratio into my definition of Controversial films. The Polar Ratio, defined as as the number of Positive Polar Ratings divided by the number of Negative Polar Ratings an individual film receives, is allowed to range between 0.4 and 2.5. In other words, one end of the polar rating spectrum cannot exceed the other by more than 2.5 times. This ensures that I am not capturing very popular, or very unpopular movies under my Controversial tag.

Below is one example of the ratings distribution a movie that fit my controversial definition, "The Passion of the Christ" (2004). Its polarized ratings made up 34.5% of all its ratings, and the ratio between positive polar and negative polar ratings was 1.62.

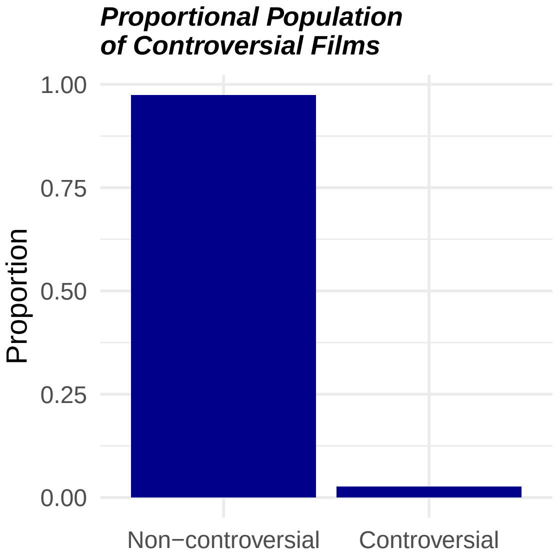


I place several restrictions on the movie dataset while developing a model for predicting controversy. The first restriction is to only consider movies that had at least 30 ratings through the MovieLens movie system. This will ensure my target class is not affected by films that were viewed, and rated, seldomly.

Additionally, I feel it is important to consider the fact that ratings motivated by controversy may be a product of the times. Issues that were considered controversial 40 years ago, may now be commonplace in today's world. A controversial movie in 1930, may be viewed as rather benign by users rating that movie in 2015.

In order to accurately capture user opinions on movies, I will only review movies that were released while the MovieLens rating system was online. From the ratings data, I observe the first review was logged on January 9, 1995. It is my determination to only review the controversiality of movies released after December 31, 1994.

With specifications and restrictions set for for my target variable, I found that 97% of movies in the dataset were considered not controversial, while only 3% of movies fit my specifications to tag as controversial. This problem is one of an imbalanced classification, for which I will make modeling corrections (to be discussed later).



Feature Selection

In order to predict controversiality of films, I will use 6 main categorical predictors, which will be used to make up over 352 binary predictors for modeling purposes. Below I provide a brief overview of the predictors:

- **Keyword Clusters**

A major source of information with this dataset is the keyword data provided by MovieLens. The keywords are words or short phrases assigned to the film that provide clues as to the content of the movie. For example, the keywords for the film "Finding Nemo", a film about a clownfish searching for his lost son include: 'Father Son Relationship', 'Harbor', 'Clownfish', and 'Protective Father'.

Assuming that the controversiality of a movie is driven by the content of the film, it is worth analyzing the keywords data to discover interesting patterns.

However across the films in this dataset, there existed nearly 20,000 associated unique keywords. To reduce this to a reasonably sized set of predictors, I employed the Word2Vec algorithm. This algorithm converts text into a 300-dimension numeric array. Words with similar meanings are mapped close together in the vector-space.[6] This allowed me to transform each keyword into a numeric cluster, then run a K-means clustering algorithm to group linguistically-similar keywords.

This process is exemplified by a word cloud of a random cluster:

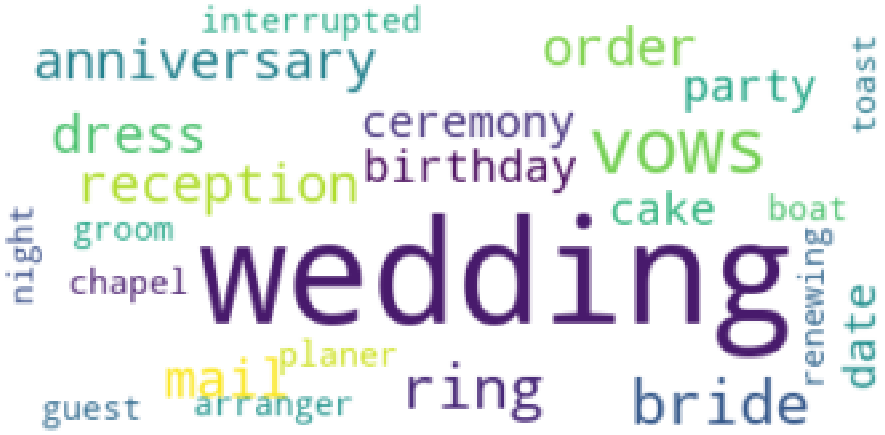


Figure 1: *Keyword Cluster #143*

After running several tests, using the elbow method and visual inspection to determine the optimal number of clusters, I decided on using 300 keyword clusters. While this did not create perfectly pure clusters, I had to balance the optimal number of clusters with the added complexity the number of predictors would entail.

Once keyword clusters were assigned, I could then generate one hot encodings for each film, creating a binary columns indicating that a movie was associated with keywords in to a certain cluster.

- **Genre Classes**

IMDb metadata data included classifications of movies across 21 different genres such as "Action/Adventure" or "Romance". Individual movies could be assigned multiple genres. Genres features were coded with one hot encoding, creating 21 columns of binary data indicating certain genre schemes for individual films

- **Original Language**

IMDb metadata also reported the original language of the film. While most films were English-language, there were many of differing languages. Language could be a proxy for country of

[6] A Beginner's Guide to Word2Vec and Neural Word Embeddings. (n.d.). Retrieved May 13, 2020, from https://pathmind.com/wiki/word2vec

production, and perhaps a tell whether a certain country's cinema was more prone to producing more controversial films. I used one hot encoding to create binary columns to indicate a film's language.

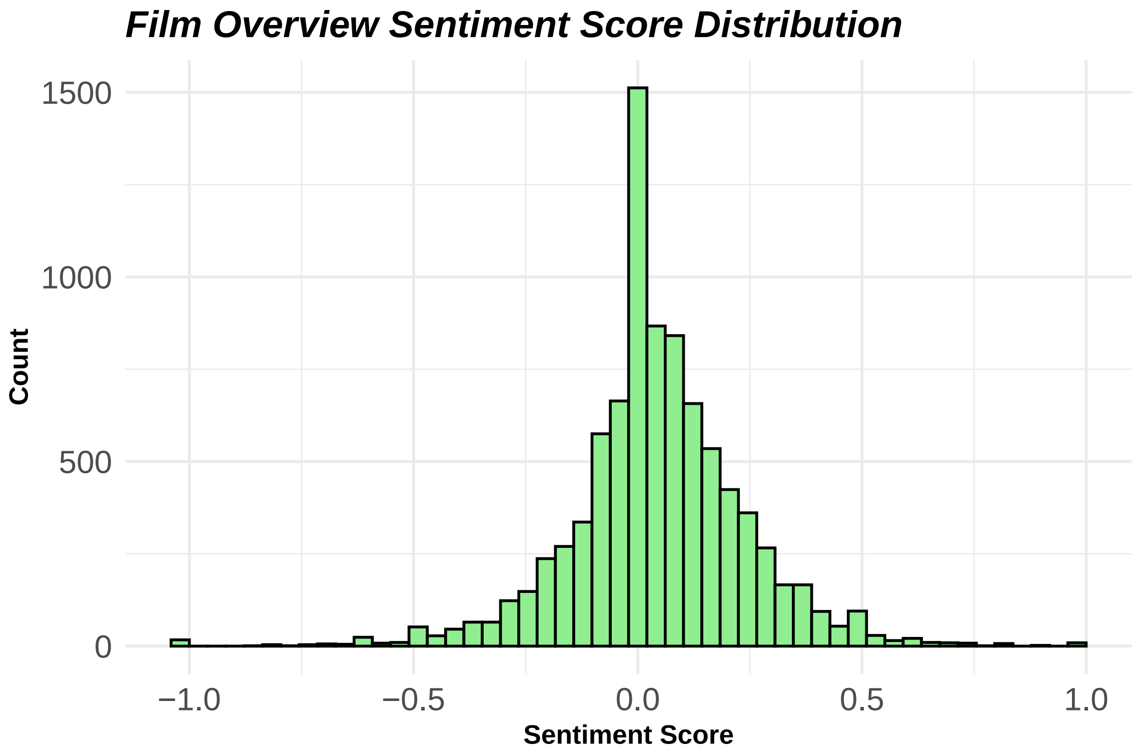
- **Sentiment Scores**

IMDb metadata also contained columns with a brief full-sentence summary of the film's plot, referred to as the "Overview". Additionally, the metadata contained a "tagline", or a short slogan used for the film's marketing. To transform this text data into a useful metric, I analyzed these features on the basis of sentiment. To analyze sentiment I used Python's TextBlob module, which processes text and assigns it a score of -1 if the text displays negative emotions, and +1 if the text is identified with positive emotions.

For the film's overview I calculated sentiment scores for each sentence comprising the summary, and then averaged the scores to create an overall sentiment score for the film's overview. For the film's tagline I processed the single sentence with the TextBlob module to create a single score for the film's tagline.

This information could be useful as it may indicate the tone of a film, where I expect dark, violent, or depressing films to have low scores, while more positive films to have higher scores - again providing clues into the content of each film.

The below histogram shows the distribution of polarity scores of the each film's overview.



- **Production Companies**

Finally, I analyzed the production companies involved in the making the films. From the IMDb metadata I observed that films had multiple production companies involved in a single film, with over 300 different production companies appearing in the data.

To transform these feature into a more manageable set of predictors, I examined the number of films each production company was involved in, and then classified those companies into four different sizes: Large, Medium, Small, and Single.

Single studios are those that were involved with in production of only a single movie in the database, and made up over 90% of the studios. Small studios were those that were involved in the production of 10 films or fewer. Medium studios were involved in the production of 60 films or fewer, and Large studios made up the remaining studios. Removing any films that did not have MovieLens keywords from the dataset, I am left with 7,739 movies and 351 features with which to predict controversiality. This dataset will be split 70/30 into a training and validation set, respectively.

As mentioned earlier, because I am dealing with an unbalance classification problem I also take the step to create an oversampled training set. This training set is created by sampling the current 70% training set, with replacement, until the minority class (Controversial films) comprise at least 40% of the training set.

Predictive Models

In targeting film Controversiality, I used three different classification methods: Logistic Regression, Gradient Boosted Random Forest, and a Neural Network classifier.

Logistic Regression

At the 50% cutoff level on the training set, the Confusion Matrix for reports an accuracy of 96.17%, with a sensitivity of 0.01 and specificity of 0.99. These scores were expected of an unbalanced classification task. The Confusion Matrix, below, shows that the classifier was stringent - only predicting the minority class in less than 1% of instances.

```
##            actual
## predictions    0    1
##           0 2232   70
##           1   19    1
```

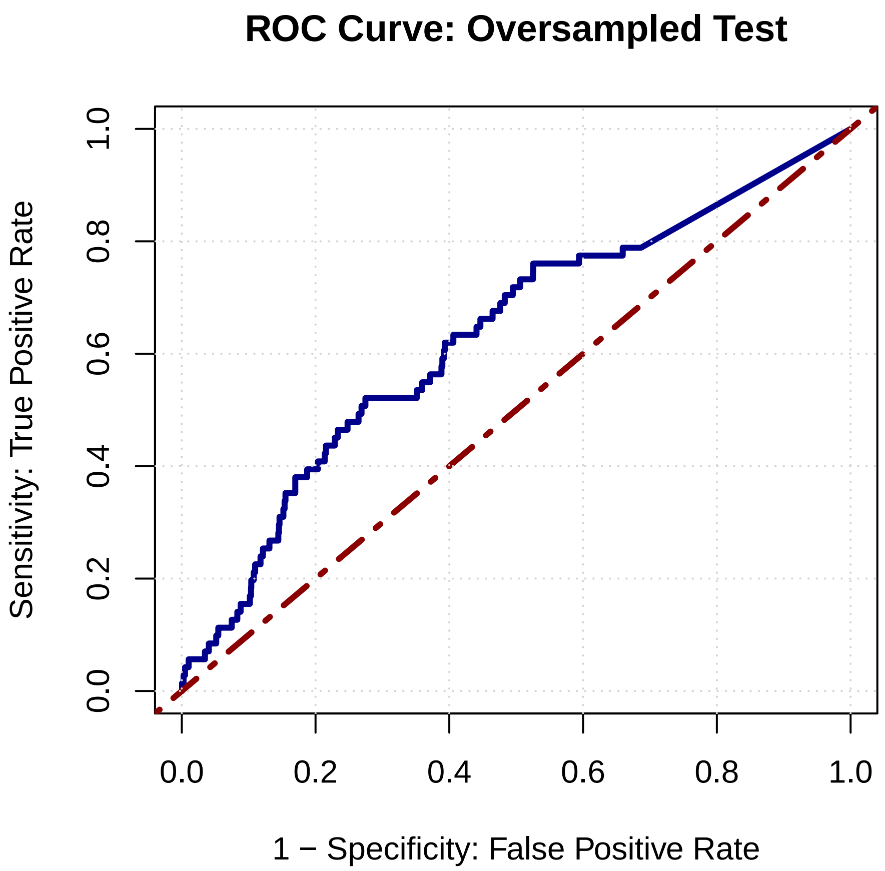
Reducing the threshold probability to 10%, increases the willingness of the classifier to assign to the minority class, increasing sensitivity increases to 0.11, but at the expense of specificity (0.95) and accuracy (92%).

```
##            actual
## predictions    0    1
##           0 2133   63
##           1  118    8
```

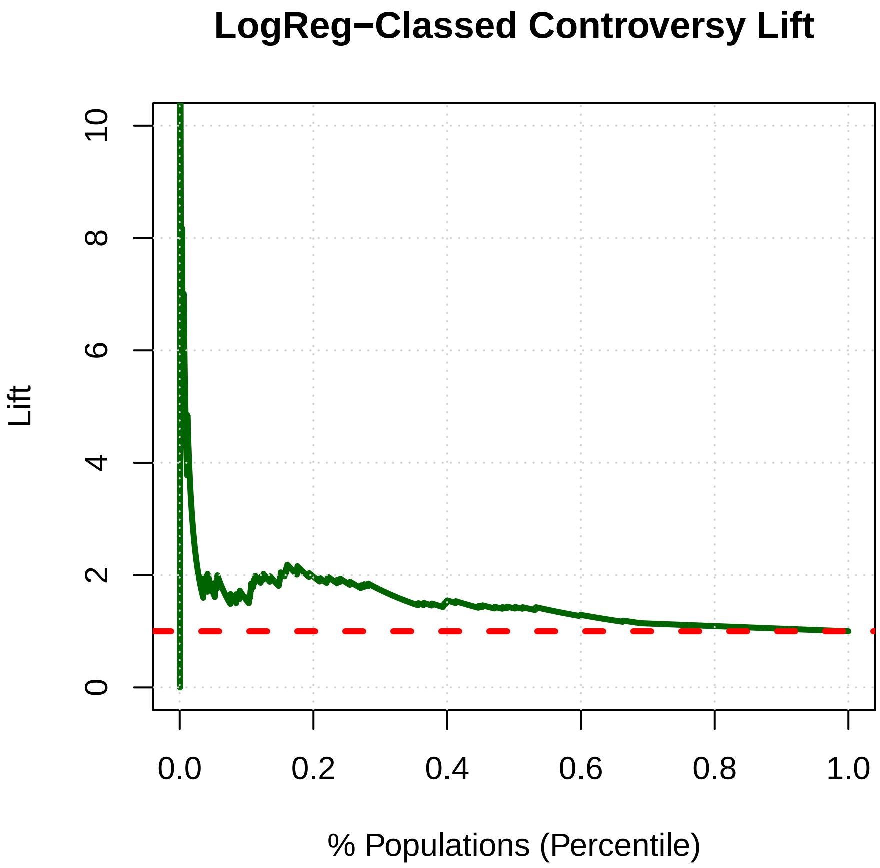
To attempt to remedy the issue of the stringent classification model, I use the logistic regression model trained on the oversampled data to redo my predictions. At the 50% cutoff level, the model scored an accuracy of 87% on at a 50% cutoff level, with a corresponding sensitivity of 0.2.

```
##            actual
## predictions    0    1
##           0 2017   57
##           1  234   14
```

The ROC curve for this logistic model is shown below, with an area under the curve of 0.63.



Translating the model's scores into a lift chart shows the limited value of the model, with its lift declining below 2.0 at very early percentiles. .



Gradient Boosted Machine

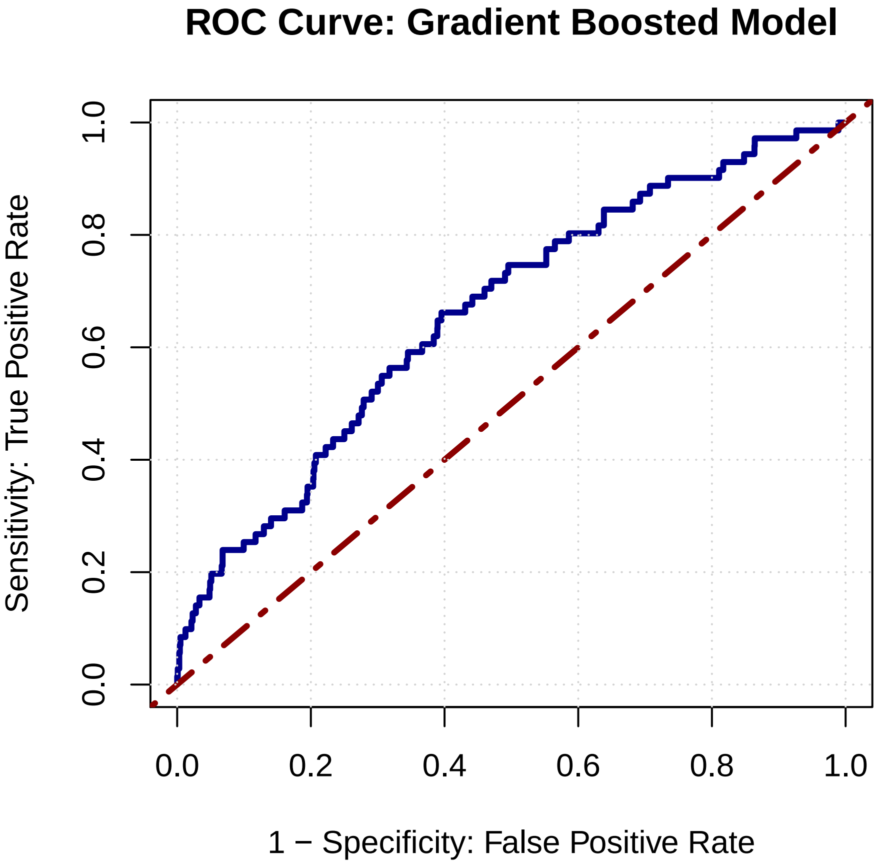
With minor success using logistic regression, I try a gradient boosted machine for classification purposes. Based on the improved sensitivity performance observed in the logistic regression, and to avoid the trap of a majority-class-only model, I will be using the oversampled dataset as the model's training set.

I trained and tuned the model using caret's GBM method, across interaction depths ranging from 5 to 15, tree count ranging from 100 to 1000, and shrinkage values of .01 and 0.1. The best model from this training had an interaction depth of 15, and shrink of 0.1, using 1000 trees.

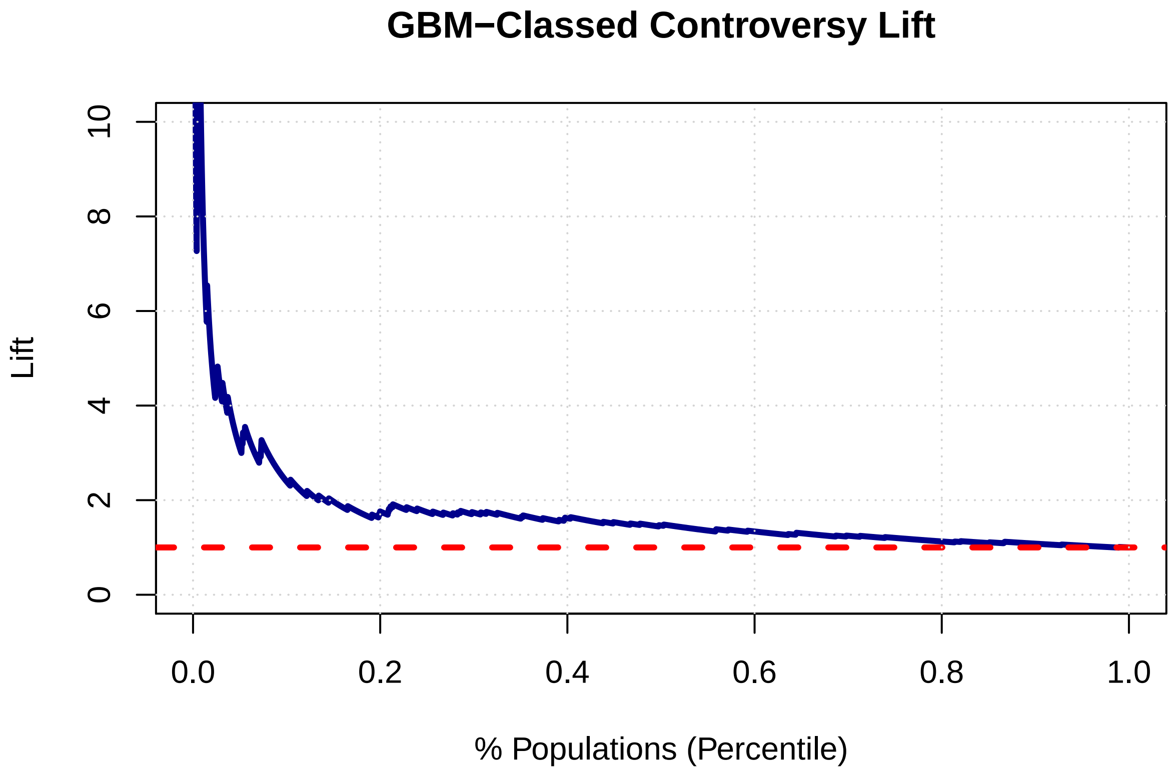
The confusion matrix below gives an accuracy of 96.34%, with a sensitivity of 0.08 and specificity of 0.96 at the 10% cutoff level.

```
##            actual
## predictions    0    1
##           0 2228   65
##           1   23    6
```

At a 10% cutoff level, the ROC curve is shown below, reporting an area under the curve of 0.65 - a slight improvement from the logistic regression.



Lift curve for GBM model is shown below, with much of the lift falling before the tenth percentile. However, the lift for this GBM model does appear to be slightly superior to that of the logistic regression.



I also identified the top influential features in the model:

- Overview Polarity
- Tagline Polarity
- Large Studio Involvement
- Thriller, Horror Genres
- Keyword Clusters 201 and 45

It could be expected that Thriller and Horror genres are among the most controversial, as these films often try to frighten viewers with shocking or violent scenes - which may be controversial depending on a viewers' appetite for that material. Keyword Cluster #45 is heavily themed with religious keywords. Traditionally, religion has elicited strong emotional reactions of people throughout history (e.g. the Crusades), and it is not unexpected that films dealing with religious topics could provoke the same type of polarizing responses.

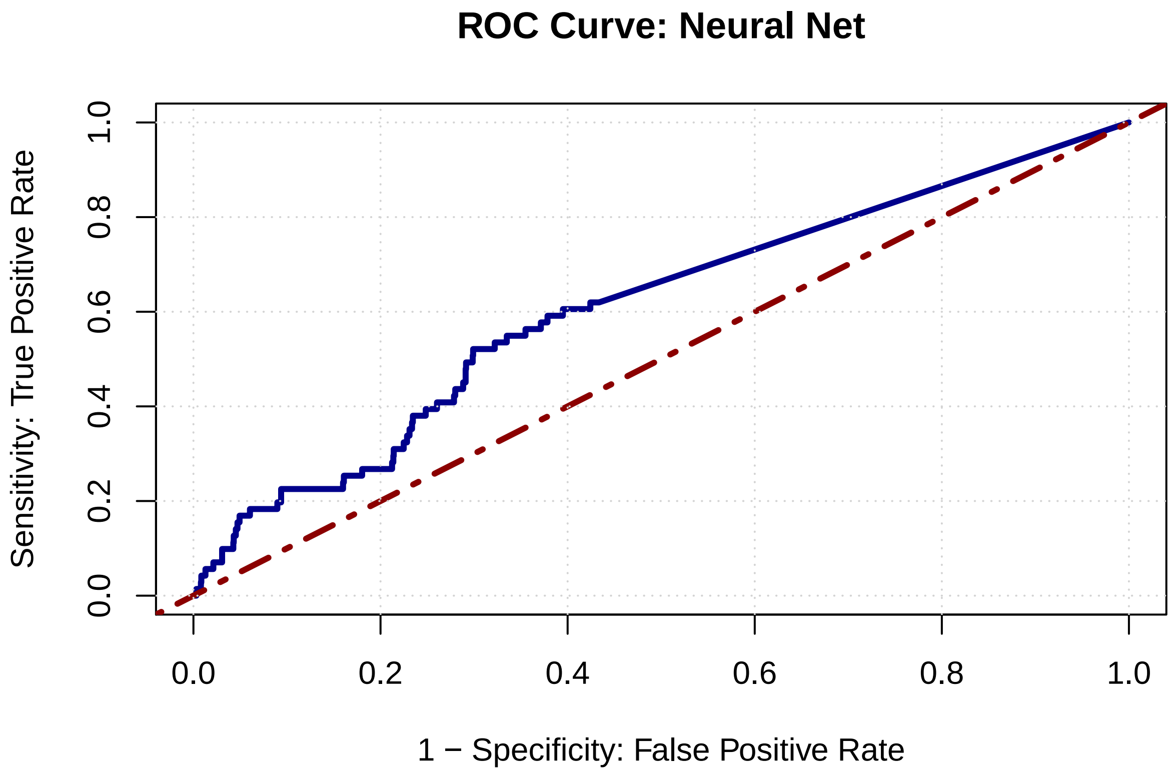
Neural Net

Finally I developed an artificial intelligence solution to assign a flag of Controversiality via the nnet package. This model was developed through ten-fold cross validation, focusing on sensitivity. From the testing I found that the neural network with more hidden layers greatly reduced the number of false positives, with only a minor tradeoff in sensitivity. I eventually tuned the model to reach a neural network with a size of 15, and a decay of 0.01, trained on the oversampled training set.

The neural network with 15 hidden layers reported an accuracy of 96%, with a sensitivity of just 0.06. The confusion matrix below shows the model was stringent in assigning minority class labels.

```
##            acutal
## predictions    0    1
##           0 2209   67
##           1   42    4
```

The ROC Curve for this model is shown below with an area under the curve of 0.63 - a similar performance to my previous models



For the neural network, the top influential variables were:

- Tagline and Overview Polarity
- Romance and Comedy genres
- Single and Small Studio Involvement
- Keyword Cluster 223 and 241

I note, with interest, that the GBM model determined ‘Thriller’ and ‘Horror’ genres to be the most influential, while the neural net determined that ‘Romance’ and ‘Comedy’ genres had the greatest influence. These genres may be viewed as almost opposites among the available genres. Further keyword cluster 241, seems to encompass keywords that would appear with horror films, such as “gore”, “slasher”, and “dystopia”.

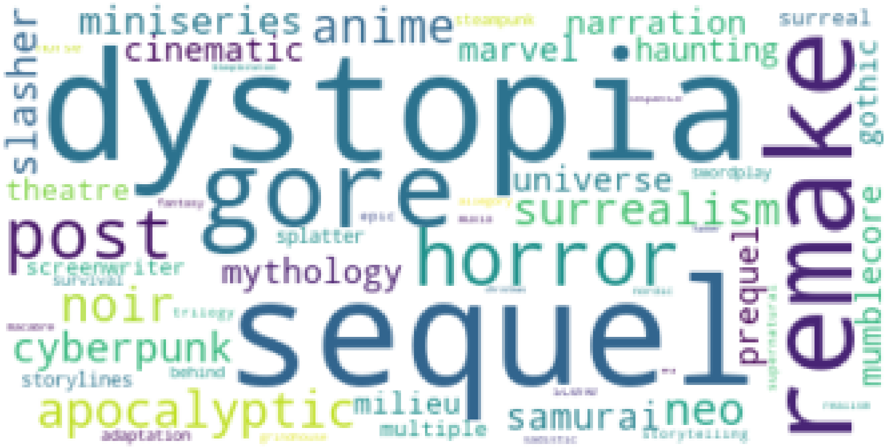
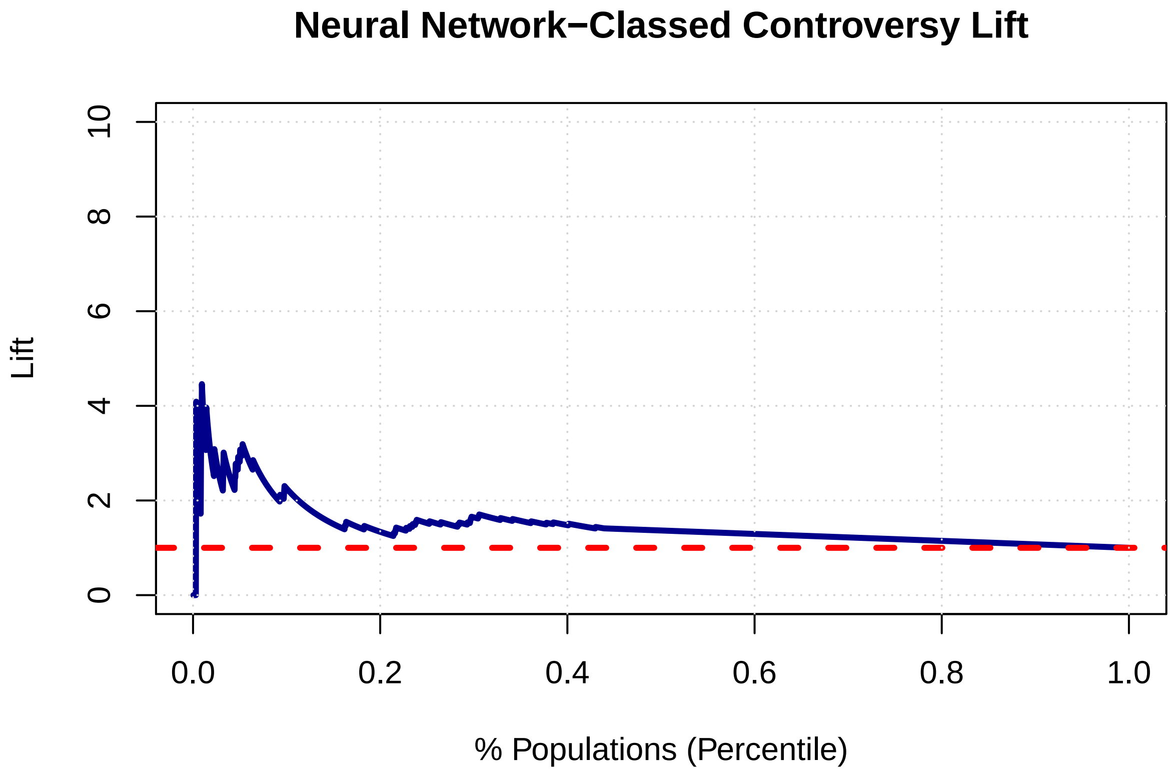


Figure 2: *Keyword Cluster #241*

Trying to make sense of tagline and overview polarity as influential features is not straightforward. One possible explanation may be that strongly negative or positive contents of a film are reflected through the overview, and the marketing for the film, and those strong sentiments of the contents are channeled in the audience's reactions to film. However, this is purely conjecture, and more study would be required to determine the actual cause.

The Lift chart for the neural network is reminiscent of that for the logistic model, with steep declines in early percentiles.



Conclusion

While the GBM model performed the best, achieving the highest area under the curve, none of the methods distinguished itself as a means to confidently predict the controversiality of films from its contents. The main issue was low sensitivity, which left the models with limited ability to identify True Positives.

This result is not necessarily surprising as controversies may often be the result of outside influence, rather than features inherent to the film. Future analyses may be better served by pulling in additional data, such as social media mentions, or counts of relevant news articles at the film's release.

Another important point to discuss in considering the fallibility of the models, is that the text analysis for keyword clusters and sentiment analysis is only as strong as the natural language processing models. These models are dependent on the linguistic idosyncracies of the base training corpus. For the keyword clustering I used a general Google News text corpus. This meant some film keywords were not able to be vectorized, as there were words in the film keyswords that did not appear the Google corpus. In future analyses, the vectorization may be better served by a corpus that is more aligned with text that would be found in films.

Finally, I believe that further steps could be taken in this analysis to begin to study the interactions between the various features. For example, more exploratory analysis could be performed to search for similarities among movies are associated with the same keyword clusters. This may inform better features to use to represent a film's contents.